Smart Anomaly Prediction in Nonstationary CT Colonography Screening

Yuichi Motai, *Senior Member, IEEE*, Dingkun Ma, and Hiroyuki Yoshida, *Member, IEEE*

***Abstract*—To enhance the quality of economically efficient healthcare, we propose a preventive planning service for next-generation screening based on a longitudinal prediction. This newly proposed framework may bring important advancements in prevention by identifying the early stages of cancer, which will help in further diagnoses and initial treatment planning. The preventive service may also solve the obstacles of cost and availability of scanners in screening. For nonstationary medical data, anomaly detection is the key problem in the prediction of cancer staging. To address anomaly detection in a huge stream of databases, we applied a composite kernel to the prediction of cancer staging for the first time. The proposed longitudinal analysis of composite kernels (LACK) is designed for the prediction of anomaly status and cancer stage for further diagnosis and the future likelihood of cancer stage progression. The prediction error of LACK is relatively small even if the prediction is made far ahead of time. The computation time for nonstationary learning is reduced by 33% compared with stationary learning.**

***Index Terms*—Anomaly detection, computed tomographic colonography, healthcare systems, kernel feature analysis, longitudinal prediction, nonstationary datasets.**

Nomenclature

Terminologies:

AKFA	Accelerated kernel feature analysis.
APH	Area of pixel histogram.
AUC	Area under the curve.
CAD	Computer-aided diagnosis.
KFA	Kernel feature analysis.
KPCA	Kernel principal component analysis.
LACK	Longitudinal analysis of composite kernels.
LMKL	Localized multiple kernel learning.
NRMSE	Normalized root-mean-square error.
PCA	Principal component analysis.
RBF	Radial basis function.
ROC	Receiver operating characteristic.
SVM	Support vector machine.
TPR	True positive rate.
VOI	Volumes of interest.

Formulas:

x_i	Input data
y_i	Output class label.
$k_{ij} = \langle \Phi(x_i), \Phi(x_j) \rangle$	Element of the Gram matrix.
K	Kernel Gram matrix.
$q_l(.)$	Factor function.
l	Base kernel label.
α_l	Combination coefficient.
K_l	Data-dependent kernel.
Q_l	Factor function matrix.
$Q_{l'}$	Updated Q_l matrix.
$P_l..$	Base kernel matrix.
$P_{l'..}$	Updated $P_l..$ matrix.
J	Class separability.
S_{br}	Between-class scatter matrices.
S_{wr}	Within-class scatter matrices.
r	Number of class.
λ_l	Eigenvalue.
λ_l^*	Largest eigenvalue.
ξ	Ratio of the class separability.
$K_{\text{comp}}^s(\rho)$	Composite kernel.
ρ	Composite coefficient.
s	One of four composite kernels.
$A(k_1, k_2)$	Empirical alignment between kernels k_1 and k_2.
δ	Eigenvalues.
$\hat{\rho}$	Optimum composite coefficient.
δd	Eigenvalues of the Database d.
$J'_*(\alpha'_r)$	Class separability yielded by the most dominant kernel for dataset(subsets) of database.
$J_*(\alpha_r)$	Class separability yielded by the most dominant kernel for the "entire" database.
a'_*	Combination coefficients of dominant kernel.
$\bar{a}_*$	Mean of combination coefficients of all database.

Manuscript received January 30, 2015; revised August 21, 2015, January 16, 2016, April 8, 2016, and June 13, 2016; accepted July 18, 2016. Date of publication July 27, 2016; date of current version December 6, 2016. This work was supported in part by CTSA UL1TR000058 from the National Center for Advancing Translational Sciences, Center for Clinical and Translational Research Endowment Fund of Virginia Commonwealth University, CAREER Award 1054333 from the National Science Foundation, and R01CA166816 from the National Institutes of Health. Paper no. TII-15-0165.R4 *(Corresponding author: Y. Motai.)*

Y. Motai and D. Ma are with the Department of Electrical and Computer Engineering, Virginia Commonwealth University, Richmond, VA 23284 USA (e-mail: ymotai@vcu.edu; madingkun@gmail.com).

H. Yoshida is with the Department of Radiology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114 USA (e-mail: yoshida.hiro@mgh.harvard.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TII.2016.2595399

I. Introduction

ANOMALY detection has long been an obstacle in the detection of cancer. The current state-of-the-art diagnosis techniques still fail to identify the important transitional cases over longitudinal patient datasets [1]. Anomaly detection in

nonstationary medical data is critical for the diagnoses of changes in staging, and thus for decision making in the treatment of cancer [2]. In CT colonography (also known as virtual colonoscopy), for example, it has long been desired to improve performance of CAD [9] in differentiating polyps from false-positive (FP) detections based on longitudinal data [6]. The diagnostic accuracy of the CT colonography can be improved, if the proposed anomaly detection in CAD can determine how much or when the next CT colonography is needed to maintain a high diagnostic accuracy [3]. The main problem pertinent to nonstationary CT colonography datasets is how to detect and predict anomaly cases or changes in the data streams over time. Thus, the purpose of this study is to detect anomalous cases based on a clinical longitudinal CT colonography datasets.

As for terminology, we call the data "normal" when they regenerate the existing data class. Conversely, the data are "anomalous" when the class is degenerated with nonstationary data. In detecting whether a polyp is normal or anomalous using the proposed algorithms, it is important to understand the data acquired "over a long period of time" can often be highly diverse and suffer from numerical errors, and each dataset is unique in nature. Therefore, obtaining a clear distinction between "normal" and "anomalous" large nonstationary datasets is a highly challenging task [7], [10]. Thus, the specific aims for anomaly detection in this study are 1) to introduce the criteria for classifying polyps as normal or anomalous, and 2) to predict the consistent progress of the anomaly in the dynamic nonstationary environment. Nonstationary cancer data in real world are highly nonlinear and unbalanced, and underlying distribution changes over time [8]. We seek to improve automated classification and prediction performance of the CAD systems in a more realistic setting, in which CT colonography datasets of new patients are added to a preestablished database on a regular basis [5].

To address the problems of nonstationary, nonlinear, and unbalanced datasets, various methods on pattern classification [14] have been proposed. Historically, time-varying pattern recognition techniques such as time-varying perceptrons, were proposed [25]. Detecting and adapting classifiers to the changes in the underlying data distributions are active areas of research. A few examples of these methods are multiple kernel learning (MKL) and incremental PCA [10]. Effective applications of online learning to nonlinear space have been investigated using kernel-based methods [11], [12]. However, the drawback of these methods is the lack of "prediction of future event/steam." If the batch of data in the underlying dynamic distribution changes over time, the time-variant pattern analysis is preferred.

In this study, we propose a new nonstationary learning technique called LACK, which substantially extends the following stationary learning techniques: AKFA [13] and principal composite kernel feature analysis (PC-KFA) [19]. The key technique in our approach is to construct a composite kernel from data-dependent kernels [19]. The composite kernel modifies itself to optimize the choice of the kernel. To reduce the weakness of traditional composite kernel methods, we develop nonstationary data associations using LACK to correspond with anomaly datasets of nonstationary cases.

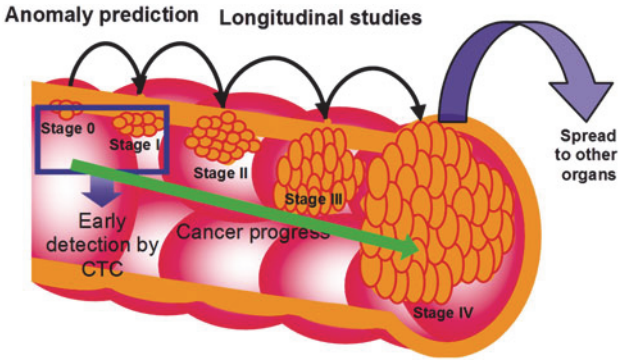

Fig. 1. Conceptual illustration of colorectal cancer staging. The degree of cancer progress is indexed by stages 0–IV, corresponding to the size of the tumor from 5 mm to a few centimeters [2].

The contribution of this paper is that the proposed LACK method yields high detection and prediction accuracy in the change of cancer staging. The proposed LACK algorithm is applied to the improvement of the performance of a CAD scheme in the detection of polyps from the clinical CT colonography database. The performance evaluation includes the accuracies of the detection and prediction of anomalies and the overall computing cost, in order to determine whether LACK can be a light-weighted module of nonstationary prediction.

The remainder of this paper is organized as follows. In Section II, we extend our stationary kernel method to nonstationary kernel method to accommodate normal and anomalous cases. In Section III, we describe how the LACK method detects and predicts the status of "normal" or "anomalous" from the incoming dataset. The experimental results of longitudinal prediction are described in Section IV, and Section V presents the conclusions.

II. Longitudinal Nonstationary Data With Anomaly/Normal Detection

In this section, we first define nonstationary data in comparison with stationary data for the target application in Section II-A. Then, we explore class separability based on the nonstationary longitudinal data in Section II-B and composite kernel for nonstationary data in Section II-C.

A. Stationary Versus Nonstationary Data for Kernel Learning

The success of the kernel learning method depends on the successful construction of the Gram matrix [19]. More specifically, the key learning element is to design a problem-specific kernel function. In the target application of longitudinal analysis of nonstationary CT colonography data, Table I lists the primary distinction between stationary and nonstationary data.

Nonstationary data, the third column in Table I, highlights a longitudinal extension of stationary data analysis for the predictive cancer staging. The nonstationary data are flexible, with varying size of data, used for iterative updates during the prediction stage. Even if the size of data increases over the examination period, the nonstationary data should be handled by appropriately extracting dominant features in a custom representation.

TABLE I
COMPARISON BETWEEN STATIONARY AND NONSTATIONARY DATA

Characteristics	Stationary	Nonstationary
Size	Fixed	Dynamic
Representation	Noniteration	Iteratively update
merit	Open loop, speed	Closed loop, flexibility
CT colonography	Existing application	Longitudinal extension

A new criterion for training the nonstationary longitudinal data has been desired for a solid mathematical representation using kernel methods. For efficient feature analysis of nonstationary dataset, extraction of the salient features of polyps is essential because of the incremental size and nonlinear nature of the polyp datasets. The problem is how to select such an ideal nonlinear positive-definite kernel operator: $k : R^d \times R^d \rightarrow R$ of an integral operator, where d is the data dimension of feature space. Some of the commonly used kernels are linear, polynomial, Gaussian RBF, and Laplace RBF.

Kernel selection substantially depends on the nonstationary data characteristics. For example, the linear kernel is important for large sparse data vectors, and it implements the simplest of all kernels, whereas the Gaussian and Laplace RBFs are general-purpose kernels used when prior knowledge about data is not available. The Gaussian kernel avoids the sparse distribution caused by the high-degree polynomial kernel in large feature space. The nonstationary data may include a wide variety of characteristics in the incoming data stream. For these reasons, we are developing advanced composite methods with data-dependent kernel customization. Our new kernel feature analysis for nonstationary longitudinal CT colonography is an extension of the following methods: KPCA or AKFA [13], and PC-KFA [19].

Let the data-dependent kernel matrix K_r for $r = 1, 2, 3, 4$ correspond to

$$k(x_i, x_j) : K_r = [\langle q_r(x_i)^T q_r(x_j)\rangle p_r(x_i, x_j)]_{n\times n} \quad (1)$$

where $\{x_i, x_j\}(i, j = 1, 2, \ldots, n)$ are n training samples of stationary and nonstationary datasets. The diagonal matrix of factor elements $\{q_r(x_1), q_r(x_2), \ldots q_r(x_n)\}$ is calculated to a factor matrix Q_i, and the elements of $p_r(x_i, x_j)$ are chosen among the four common kernels to form P_r [14]. Now, we can express (1) as $K_r = Q_r P_r Q_r$.

A composite kernel function is defined as the weighted sum of the set of different optimized kernel functions using the PCA [13], [19]. To obtain the optimum detection accuracy, we define the composite kernel as

$$K_{\text{comp}}(\rho) = \sum_{i=1} \rho_i Q_i P_i Q_i \quad (2)$$

where the value of the composite coefficient ρ_i is a scalar value, and p is the number of kernels we intend to combine. This composite kernel matrix $K_{\text{comp}}(\rho)$ satisfies Mercer's condition [15]. The criteria for selecting the dominant kernel are to select the largest eigenvalues corresponding to the kernel Gram matrix $K_{\text{comp}}(\rho)$.

Among the incoming nonstationary datasets, there is a chance that the data may cause problems during cancer detection and prediction. To handle these potential complications of nonstationary datasets, we adopt the notion of "normal" or "anomalous" to make the algorithm robust [16]. The key idea of the detection of normal/anomaly is to allow the feature space to be updated appropriately as the training proceeds with more data being fed into the nonstationary algorithm, described in the next subsection.

B. Class Separability Based on the Nonstationary Longitudinal Data

We propose "class separability" as a measure to identify whether the nonstationary data are anomaly or normal. The class separability is expected to be maximized so that the training data can be separately clustered. The traditional measure to represent the class separability J of the training data is formulated as

$$J = \sum_r \text{tr} S_{br} \Big/ \sum_r \text{tr} S_{wr} \quad (3)$$

where S_{br} represents "between-class scatter matrices" and S_{wr} represents "within-class scatter matrices" known as Fisher scalar [24]. Suppose that the training data are grouped according to their class labels, such that the first n_1 data belong to one class, and the remaining n_2 data belong to the other class $(n_1 + n_2 = n)$. Then, the basic kernel matrix P_r can be partitioned as $P_r = [P^r_{11}, P^r_{12}; P^r_{21}, P^r_{22}]$, where $r = 1, 2, 3, 4$ represents the four kernels, and the size of the submatrices $P^r_{11}, P^r_{12}, P^r_{21}$, and P^r_{22} are $n_1 \times n_1$, $n_1 \times n_2$, $n_2 \times n_1$, and $n_2 \times n_2$ respectively. According to [18], the class separability in (3) can be expressed as

$$J(\alpha_r) = \frac{\alpha_r^T M_{0r} \alpha_r}{\alpha_r^T N_{0r} \alpha_r} \quad (4)$$

where $M_{0r} = K_0^T B_{0r} K_0$, and $N_{0r} = K_0^T W_{0r} K_0$

$$B_{0r} = \begin{bmatrix} \frac{1}{n_1} P^r_{11} & 0 \\ 0 & \frac{1}{n_2} P^r_{22} \end{bmatrix} - \frac{1}{n-1} P_r$$

$$W_{0r} = \text{diag}(p^r_{11}, p^r_{22}, \ldots, p^r_{nn}) - \begin{bmatrix} \frac{1}{n_1} P^r_{11} & 0 \\ 0 & \frac{1}{n_2} P^r_{22} \end{bmatrix}.$$

The maximized method of each class separability is derived in [19]. Let us introduce a variable ξ, which is the ratio of the class separability of the composite (both stationary and nonstationary) data and the stationary data. It can be expressed as

$$\xi = J'_*(\alpha'_r)/J_*(\alpha_r) \quad (5)$$

where $J'_*(\alpha'_r)$ denotes the class separability yielded by the most dominant kernel (which is chosen from the four different kernels) for the composite data (i.e. new incoming data and the previous stationary data), and $J_*(\alpha_r)$ is the class separability yielded by the most dominant kernel for stationary data. Because of the properties of class separability, (5) can be rewritten as $\xi = \lambda'_*/\lambda_*$, where λ'_* corresponds to the most dominant eigenvalue of composite data (both stationary and nonstationary), and λ_* is the most dominant eigenvalue of the four different kernels for the stationary data. If ξ is less than a threshold value 1.0,

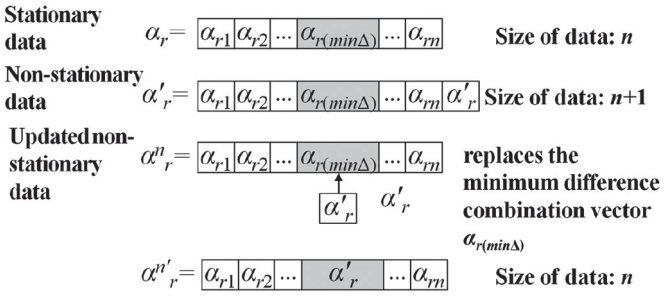

Fig. 2. Update of combination vector for Case 1: Minor Anomaly. The trivial rows and columns of the previous Gram matrix with those of the new data are replaced as follows $\alpha_{r(\min\Delta)} \leftarrow \alpha'_r$.

then the incoming nonstationary data are anomaly; otherwise, these are normal.

1) *Normal nonstationary data:* In the case of normal data, we do not update the Gram matrix. Instead, we discard all the data that are normal. Hence, the updated Gram matrix can be given as $K_r^{n'} = Q_r P_r Q_r$, where n' is the number of nonstationary incoming data.
2) *Anomaly nonstationary data:* If the new batch of incoming data is anomaly, we update our feature space based on the anomaly degree in the relationship between η and ξ through empirical experimental setting. The second threshold value η determines whether the kernel update is *Case 1: Minor Anomaly*, or *Case 2: Significant Anomaly*. The procedures outlined in these two cases of anomaly help to refine the kernel reconstruction by adjusting the trajectories of the incoming nonstationary datasets.

The class separability of the most dominant kernel for the new data is directly dependent on both the maximum combination coefficient of four different kernels and the maximum eigenvalue λ'_*. Let us denote α_* as the mean of combination coefficients prospectively, and α'_* for the most dominant kernel among the four kernels available. Using these values, we update the Gram Matrix for eigenrepresentation.

Case 1: Minor Anomaly: Under the case of minor anomaly if $\eta \leq \xi\ (\leq 1)$, the dimensions of the Gram matrix remain constant. We search for small trivial elements in the Gram matrix and then replace the trivial rows and columns of the previous Gram matrix with those of the new data by calculating the minimum difference vectors $\min\Delta$ as the trivial one. We compare and replace the combination coefficient values of stationary data with the combination coefficient of the new incoming data, as shown in Fig. 2. This process is repeated for all the kernel matrices. The input matrices P' and Q' are updated by removing the row and column corresponding to the index of $\alpha_r \min\Delta$ and replacing it with the row and column corresponding to the index α'_r. Then the new input matrix $P_r^{n'}$ can be written as $P_r^{n'} = [P'_r] - [P_{r(r\min\Delta,1:(n+1))}] - [P_{r(1:n,r}\min\Delta]$.

After we compute $\alpha_r^{n'}$, we can compute $Q^{n'}$ as $Q_r^n = \text{diag}(\alpha_r^n)$. Hence, in the update of minor anomaly data, the Gram matrix can be given as $K_r^{n'} = Q_r^{n'} P_r^{n'} Q_r^{n'}$.

Case 2: Significant Anomaly: When the similarity between the previous eigenvectors and the new eigenvectors is very small, e.g., $(0 \leq)\ \xi \leq \eta$, it is very important for us to retain these new data, which are highly anomalous, for efficient detection. Therefore, instead of replacing the trivial rows and columns of the previous data, we simply retain them, and thus, the size of the Gram matrix is increased by the size of the new data. In this case, since we have already calculated the new combination coefficient α'_r, input matrix P' and Q' as same as before, the kernel Gram matrix can be given as $K_r^{n'} = Q'_r P'_r Q'_r$. These two anomaly cases and normal cases are represented in Fig. 3.

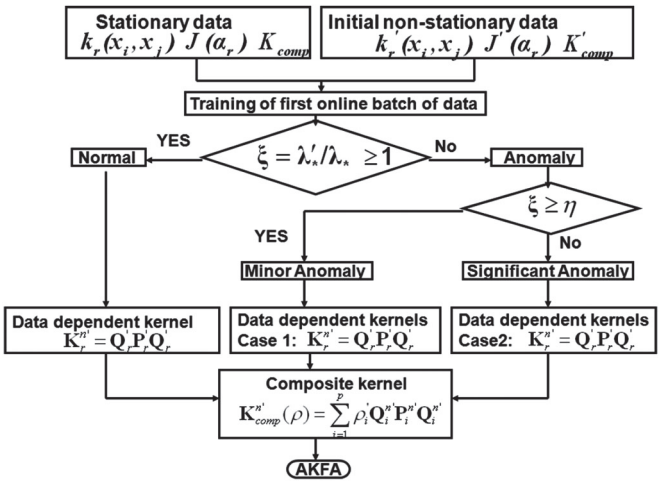

Fig. 3. Training of first nonstationary dataset using two cases to iteratively update composite kernels. Anomaly or normal detection is conducted using the proposed criterion.

C. Composite Kernel for Nonstationary Data

Once we have our kernel Gram matrices, we can now determine the composite kernel that gives us the optimum detection accuracy when the existing stationary data are incorporated with new nonstationary data. For the case of normal, minor anomaly, and significant anomaly, the identical process is executed, as shown in Fig. 3. Using Gram matrices for the (p)-most dominant kernels for the composite data (stationary plus nonstationary data), we now combine them to yield the best detection performance using AKFA [13]. We have written the composite kernel for the new composite data as: $K_{\text{comp}}^{n'}(\rho) = \sum_{i=1}^{p} \rho'_i Q_i^{n'} P_i^{n'} Q_i^{n'}$, where the composite coefficient set $\hat{\rho}'$ is the collection of combination coefficients ρ_i, and p is the number of kernels we intend to combine.

D. Comparison of Relevant Kernel Methods

Composite kernel methods are derived from the principle of theoretical approximation error bounds with respect to the kernel selection problem, e.g., Jaakkola–Haussler bound [22], radius-margin bound [23], kernel linear discriminant analysis [24], consistent dictionary learning [17], and online extreme [21]. Table II lists some comparative methods among the relevant composite kernel methods.

TABLE II
COMPARISON AMONG COMPOSITE KERNEL METHODS

Method Name	Principle	Advantages/Disadvantages
Jaakkola–Haussler bound	Leave-one-out error	Loose approximations
Radius-margin bound	Gradient of bound	Optimal parameters
Kernel linear discriminant analysis	Nonliner kernel trick	Complicated discriminant
Consistent dictionary learning	Sparse-code matrix	Adaptive label prediction
Online extreme	Recursive least-squares	Adaptive filtering

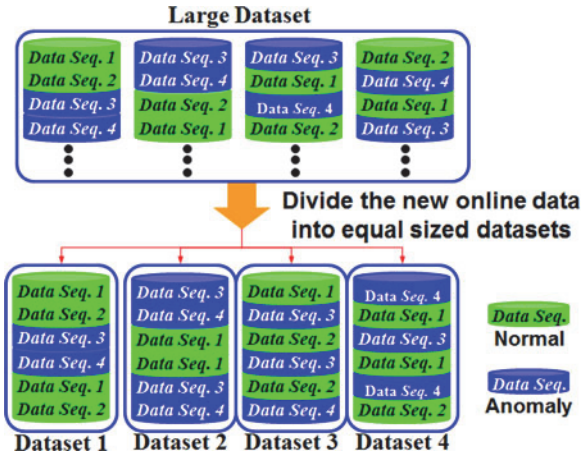

Fig. 4. Conceptual illustration of the division of a large nonstationary dataset into several small subsets of equal size. A large dataset is clustered into small subsets of datasets, for example, $1-2-3-4$, in order to make the size feasible for the time sequence.

III. LONGITUDINAL SEQUENTIAL TRAJECTORIES FOR ANOMALY DETECTION AND PREDICTION

We examine how to handle the nonstationary data over a long period of time. When the longitudinal trajectory is considered, the "long-term" sequential change is evaluated according to the criteria with the anomaly degree. The nonstationary data are suffering from the large memory size and are required to accommodate a stream or batches of data whose underlying distribution changes over time. To divide the long-term nonstationary data into smaller subsets, each of which is considered one at a time, we may treat small datasets as anomaly or normal data by checking the alignment of each dataset.

The following subsections first defines alignment factor in Section III-A by comparing existing nonstationary data so that the long-term data can be divided into subsegmented data. Then, we readdress the class separability for the long-term nonstationary data in Section III-B.

A. Nonstationary Small Chunks Datasets

Let us consider a small subset of data to find the appropriate window size for the incoming data and determine whether it is normal or anomaly data. Fig. 4 illustrates the segmentation of the incoming nonstationary data into small equal datasets l. Each dataset consists of individual data sequences in Fig. 4. All new sets of data are processed sequentially but not simultaneously.

We introduce here "alignment factor" method to determine the optimum composite coefficients that yield the best composite kernel [18]. Let our new label vector be y'_n. The alignment factor is defined as $A'(k^{n'}, y^{n'}(y^{n'})^T) = ((y^{n'})^T K^{n'} y^{n'})/(n' \| K^{n'} \|_F)$. We can determine the composite coefficient that maximizes the alignment factor as follows:

$$\begin{aligned} \hat{\rho}' &= \arg_{\rho'} \max \; A' \; \rho', k^{n'}, y^{n'}(y^{n'})^T \\ &= \arg_{\rho'} \max \; \frac{1}{n'^2} \frac{(\rho'_i)^T U^{n'} \rho'_i}{(\rho'_i)^T V^{n'} \rho'_i} \end{aligned} \tag{6}$$

where $u_i^{n'} = \langle K_i^{n'}, \; y_i^{n'}(y_i^{n'})^T \rangle$, $U_{ij}^{n'} = u_i^{n'} u_j^{n'}$, $V_{ij}^{n'} = v_{ij}^{n'} = \langle K_i^{n'}, K_j^{n'} \rangle$. Consider the new generalized Raleigh coefficient $\hat{\rho}' = ((\rho')^T U^{n'} \rho')/((\rho')^T U^{n'} \rho')$. We choose a composite coefficient vector that is associated with the maximum eigenvalue [14]. Once we know the eigenvectors, i.e., combination coefficients of the composite data (both stationary and nonstationary), we can compute q^r and hence Q'_r to find out the four Gram matrices corresponding to the four different kernels by using $K_r^{n'} = Q'_r P'_r Q'_r$. Now, the first p kernels corresponding to the first p eigenvalues (arranged in the descending order) are used in the construction of a composite kernel that yields optimum detection accuracy.

In order to determine whether or not the algorithm has been correctly trained, we make use of the alignment factor as well. Let us denote ${}^{t+1}(K^{n'}_{\text{comp}})_t$ as the update of Gram matrix from time t to $t+1$, and let ${}^{t}(K^{n'}_{\text{comp}})_0$ be the updated Gram matrix from time "0" (i.e., from the beginning till one last step) to time t. Similarly, the matrix ${}^{t+1}(y_{n'})_t$ indicates the update of the output matrix from time t to $t+1$, and ${}^{t}(y_{n'})_0$ indicates the update of output matrix till time t. The alignment factor of $[t, t+1]$ and $[0, t]$ can be given as ${}^{t+1}(A'(k^{n'}_{\text{comp}}, y_{n'}, y_n^T))_t$ and ${}^{t}(A'(k^{n'}_{\text{comp}}, y_{n'}, y_n^T))_0$, respectively:

$${}^{t+1}(A'(k^{n'}_{\text{comp}}, y_{n'}, y_n^T))_t < {}^{t}(A'(k^{n'}_{\text{comp}}, y_{n'}, y_n^T))_0. \tag{7}$$

If the variation in the alignment factor exists, the incoming nonstationary data are determined as too large, i.e., the alignment factor of the incoming data is larger than the average of the alignment factor $[0, t]$. In order to correctly train, the nonstationary data are divided into small chunks. The division stops if the new alignment factor is satisfied (7).

B. Class Separability Based on the Long-Term Nonstationary Longitudinal Data

Let us consider the new $d1$ data that have been derived from nonstationary small chunks datasets above. The input matrix at the current state, i.e., at time $(t+1)$, should be ${}^{(t+1)}(P_r^{n'})_t = [{}^{t}(P_r^{n'})_0 \; P_{d1}]$. Similarly, at time $(t+1)$, according to (4), calculate ${}^{(t+1)}(B_{0r}^{d1})_t$, ${}^{(t+1)}(W_{0r}^{d1})_t$, ${}^{(t+1)}(M_{0r}^{d1})_t$, and ${}^{(t+1)}(N_{0r}^{d1})_t$. The class separability of the composite data (data up to time t and $d1$ data) can be given as

$${}^{t+1}(J^{d1}(\alpha_r))_t = \frac{{}^{t+1}(\alpha_r^{d1})_t^T * {}^{t+1}(M_{0r}^{d1})_t * {}^{t+1}(\alpha_r^{d1})_t}{{}^{t+1}(\alpha_r^{d1})_t^T * {}^{t+1}(N_{0r}^{d1})_t * {}^{t+1}(\alpha_r^{d1})_t}. \tag{8}$$

Therefore, ${}^{(t+1)}(\xi)_t = {}^{(t+1)}(\lambda'_*)_t / {}^{t}(\lambda_*)_0$, where ${}^{(t+1)}(\lambda'_*)_t$ is the largest eigenvalue of the data (of the dominant kernel)

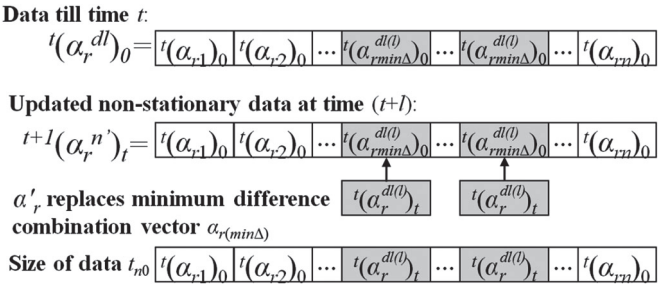

Fig. 5. Update of combination vector for Case 1: Minor Anomaly at time $t+1$ for longitudinal sequential trajectories. The data sequence with time index is evaluated for replacing the previous Gram matrix by the new time sequential datasets.

received from time t to $(t+1)$, and ${}^{t}(\lambda_*)_0$ is the largest eigenvalue of the dominant kernel calculated from time 0 to time t, i.e., till the previous step. We test whether this subdataset $d1$ is either anomalous or normal. If ${}^{(t+1)}(\xi)_t$ is greater than a new threshold value η', then the incoming data are anomalous; otherwise, these are normal.

Case 1: Minor Anomaly: If the data $d1$ are anomaly, then we decide whether these are highly or less anomaly by comparing ξ and η'. Under case 1, we find out and replace the trivial rows and columns in the dataset from the previous step and update them with the corresponding rows and columns from the new dataset $d1$. The combination coefficients in the previous step corresponding to the minimum of this difference are replaced by the corresponding vectors in the present step. This is represented in Fig. 5.

Fig. 5 shows the minor anomaly update of combination vector at time $t+1$. Similarly, update ${}^{(t+1)}(P_r^{n'})_t$ and ${}^{(t+1)}(Q_r^{n'})_t$ and compute the new Gram matrix. Now, after determining the data-dependent kernels for $d1$ data, we determine the composite kernel using (7). This case is summarized in the following steps:

Minor Anomaly Case

Step 1: Decide whether the data $d1$ are minor anomaly or not. $\eta' \leq \xi^{d1}\ (\leq 1)$.

Step 2: Find out and replace the trivial rows and columns in the dataset from the previous step and update them with the corresponding rows and columns from the new dataset $d1$ by ${}^{t+1}(\Delta_{rk}^{d1})_t = {}^{t+1}(\alpha_r^{d1})_t - {}^{t}(\alpha_{rk})_0$.

Step 3: Update ${}^{t+1}(P_r^{n'})_t$ and ${}^{t+1}(Q_r^{n'})_t$ to compute the new Gram matrix ${}^{t+1}(K_r^{n'})_t = {}^{t+1}(Q_r^{n'})_t * {}^{t+1}(P_r^{n'})_t * {}^{t+1}(Q_r^{n'})_t$.

Step 4: Determine the composite kernel as ${}^{t+1}(K_{\text{comp}}^{n'}(\rho))_t = [\sum_p \rho_i ({}^{t+1}(K_r^{n'})_t)]_{{}^{t}n_0 \times {}^{t}n_0}$

Case 2: Significant Anomaly: If the data are significant anomaly, we append the data so as to preserve the diverse information that may be used to classify the next incoming chunk of nonstationary data, i.e., $d2$ as either anomaly or normal. The data-dependent kernel for highly anomaly $d2$ data can now give the matrices with the increased size as

$$ {}^{t+1}(K_r^{n'})_t = {}^{t+1}(Q_r^{d2})_t * {}^{t+1}(P_r^{d2})_t * {}^{t+1}(Q_r^{d2})_t. \tag{9}$$

The remainder of the entire algorithm is summarized in the following steps.

Significant Anomaly Case

Step 1: Append the data to preserve the diverse information to classify the next incoming chunk of online data. Compute ${}^{t+1}(\alpha_r^{n'})_t = {}^{t+1}(\alpha_r^{d2})_t^{t+1}(Q_r^{n'})_t = \text{diag}({}^{t+1}(\alpha_r^{n'})_t) = {}^{t+1}(Q_r^{d2})_t, {}^{t+1}(P_r^{n'})_t = {}^{t+1}(P_r^{d2})_t$.

Step 2: Update the data-dependent kernel as ${}^{t+1}(K_r^{n'})_t = {}^{t+1}(Q_r^{d2})_t * {}^{t+1}(P_r^{d2})_t * {}^{t+1}(Q_r^{d2})_t$.

Step 3: Find composite kernel that yields optimum detection accuracy as ${}^{t+1}(K_{\text{comp}}^{n'}(\rho))_t = [\sum_p \rho_i ({}^{t+1}(K_r^{n'})_t)]_{({}^{t}n_0+l)\times({}^{t}n_0+l)}$.

C. Anomaly Prediction of Long-Time Sequential Trajectories

K-step prediction uses the existing data sequence to predict the next k-step value using the ARMA model, in which only the output estimates are used, and the uncertainty induced by each successive prediction is not accounted for. In order to avoid this limitation, we propose the framework of "predictive LACK" by extending the composite kernels for the k-step in ahead of time index, i.e., $(t+k)$, instead of one step of $(t+1)$.

The anomaly prediction in a clinical application is for cancer staging, which is given by a number ranging from 0 to IV, with IV having more progression, as shown in Fig. 1. The cancer stage often takes into account the size of a tumor; however, several other factors are also concerned with the stage of the cancer, such as whether it has invaded adjacent organs, how many lymph nodes it has metastasized to, and whether it has spread into distant organs. Staging of cancer is the most important predictor of survival, and cancer treatment is primarily determined by staging.

The stage of a cancer is used for the quantitative values for prediction using LACK. We extend the anomaly degree evaluated by the kernel Gram matrix into the cancer staging for the prediction framework that describes whether it has spread to distant organs. LACK adapts to the staging of cancer, the most important predictor of survival, so that cancer treatment planning can be determined by the predicted stage of cancer. The proposed framework, called predictive LACK, consists of the following five steps:

Predictive LACK:

Step 1: Regroup the datasets for applying stationary composite kernel matrix algorithm:

$k_r(x_i, x_j) = q_r(x_i) q_r(x_j) p_r(x_i, x_j)$,
$K_{\text{comp}}(\rho) = \sum_{i=1}^{p} \rho_i Q_i P_i Q_i$, and
$J = (tr(\sum_r S_{br}))(tr(\sum_r S_{wr}))$.

Step 2: Apply alignment factors to determine whether to divide datasets into small subsets:

${}^{t+1}(A'(k_{\text{comp}}^{n'}, y_{n'}, y_n^T))_t < {}^{t}(A'(k_{\text{comp}}^{n'}, y_{n'}, y_n^T))_0$,
${}^{t+1}(J^{dl}(\alpha_r))_t = \frac{{}^{(t+1)}(\alpha_r^{dl})_t^T * {}^{(t+1)}(M_{0r}^{dl})_t * {}^{(t+1)}(\alpha_r^{dl})_t}{(t+1)(\alpha_r^{dl})_t^T * {}^{(t+1)}(N_{0r}^{dl})_t * {}^{(t+1)}(\alpha_r^{dl})_t}$.

Step 3: Calculate LACK for sequences of time horizontal window, starting from 1 to k:

${}^{t+1}(K_r^{n'})_t = {}^{t}(Q_r^{n'})_0 * {}^{t}(P_r^{n'})_0 * {}^{t}(Q_r^{n'})_0$.

Step 4: Compute kernel matrix reconstruction accuracy using $MErr = (1/n) \sum_{i=l+1}^{n} \lambda_i$, with composite kernel matrix to convert the k-composited matrix value to the synthesized measurement value.

Step 5: Find out the cancer stage corresponding to the synthesized measurement data.

IV. Anomaly Detection and Prediction Results

The performance of the anomaly detection method in Section II was evaluated using CT colonography image datasets of colonic polyps that consisted of true positives (TPs) and FPs detected by our CAD system [5]. We evaluate Cancer Datasets in Sections IV-A, anomaly detection for the nonstationary data in Section IV-B, time horizontal prediction for risk factor analysis of anomaly long-time sequential trajectories in Section IV-C, and computational time for complexity evaluation in Section IV-D.

A. Cancer Datasets

The retrospective data collection and analyses carried out in this study were approved by the institutional review board at our institution. All patient records/information and image data were anonymized and deidentified prior to the analyses in this study. For the stationary experiment, we collected CT colonography data of 146 patients who had undergone a standard cathartic bowel-cleansing regimen for CT colonoscopy. Each patient was scanned in both supine and prone positions, resulting in a total of 292 CT data. The ground truth (locations and sizes of true polyps in CT colonography images) for the evaluation was established by expert radiologists with reading experience on >500 cases reference to colonoscopy and histology reports. The VOIs representing each polyp candidate were obtained as follows: The CAD scheme provided a segmented region for each candidate. The center of a VOI was placed at the center of mass of the region. The size of the VOI was chosen so that the entire region of the polyp was covered. Resampling was carried out using VOIs with dimensions of $12 \times 12 \times 12$ voxels to build Stationary Set1, which consisted of 29 true polyps and 101 FPs. For the rest of the datasets, the VOI was resampled to $16 \times 16 \times 16$ voxels. The VOIs computed were Stationary Set1 (29 TPs and 101 FPs), Set2 (54 TPs and 660 FPs), Set3 (16 TPs and 925 FPs), and Set4 (11 TPs and 2250 FPs).

For a nonstationary experiment, we used a large dataset that had a total of 3749 CT images. These data were acquired by helical single-slice or multislice CT scanners, with collimation of 0.5 mm, a reconstruction interval of 0.5 mm, X-ray tube currents of $50-260$ mA, and voltages of $120-140$ kVp. The in-plane pixel size was 0.5 mm, the CT image matrix size was 512×512, and these images were used to form one dataset. As shown in Table III, we divided the datasets into four groups. There were 368 normal cases and 54 abnormal cases with colonoscopy-confirmed polyps.

Table III shows the arrangement of stationary training and testing sets: Stationary Set1, Set2, Set3, and Set4, as well as the nonstationary training and anomaly test sets: Nonstationary Set1, Set2, and Set3. Instead of using the cross validation, we randomly divided the entire datasets into a training and a test set. We kept each testing set for each dataset for the entire experiment.

TABLE III
Arrangement of Datasets

Datasets	No. of Vectors in Training Set		Total	No. of Vectors in Testing Set		Total
	TP	FP		TP	FP	
Stationary Set1	21	69	90	8	32	40
Stationary Set2	38	360	398	16	300	316
Stationary Set3	10	500	510	6	425	431
Stationary Set4	7	1050	1057	4	1200	1204
Nonstationary Set1	15	403	418	19	500	519
Nonstationary Set2	20	503	423	28	600	628
Nonstationary Set3	25	706	731	29	900	929

TABLE IV
Value of $\hat{\rho}$ for Each Composite Kernel

Datasets	Two Dominant Kernels	Linear Combination	Reconstruction Error %
Stationary Set1	RBF and Laplace	$\rho 1 = 0.98, \rho 2 = 0.14$	1.00
Stationary Set2	RBF and Polynomial	$\rho 1 = 0.72, \rho 2 = 0.25$	9.64
Stationary Set3	RBF and Linear	$\rho 1 = 0.98, \rho 2 = 0.23$	6.25
Stationary Set4	RBF and Polynomial	$\rho 1 = 0.91, \rho 2 = 0.18$	14.03

For evaluation of the performance in the detection of cancer (TP and FP rates), we used the method proposed in Section II to create four different data-dependent kernels, selected the kernel that best fit the data, and gave optimum detection accuracy for the stationary data. We determined the optimum kernel depending on the eigenvalue that yielded maximum separability. Based on the order of eigenvalues, we selected the two largest kernels to form the composite kernel. Taking Dataset1 as an example, we combined RBF and Laplace to form the composite kernel. We observed that each database had different combinations of composite kernels. The composite coefficients for the two most dominant kernels are listed in Table IV.

Table IV shows how the two kernel functions are combined based on the composite coefficients. These composite coefficients were obtained using (2) in Section II. Among all of the datasets, the most dominant kernel was the RBF kernel, whereas the second most dominant kernel kept varying. As a result, the contribution of the RBF kernel was higher than that of the other kernels in forming a composite kernel. The reconstruction error is also shown in Table IV by calculating $E_{rri} = \|\Phi_i - \Phi_i'\|^2$. Here, the reconstruction error represents the difference between the raw data and the value yielded by the linear combination of two dominant composite kernels with linear combination. A small reconstruction error indicates that the two dominant kernels represent the original datasets well. Fig. 6 shows the ROC curves for the stationary datasets. As shown in Table IV and Fig. 6, a good classification performance was obtained over all of the four stationary datasets. In particular, the AUCs for four datasets were improved by 25% compared with an earlier study [13], [19].

Table V shows that the proposed LACK had a comparable classification accuracy for the four datasets. All the results indicate that the proposed LACK was very competitive with other

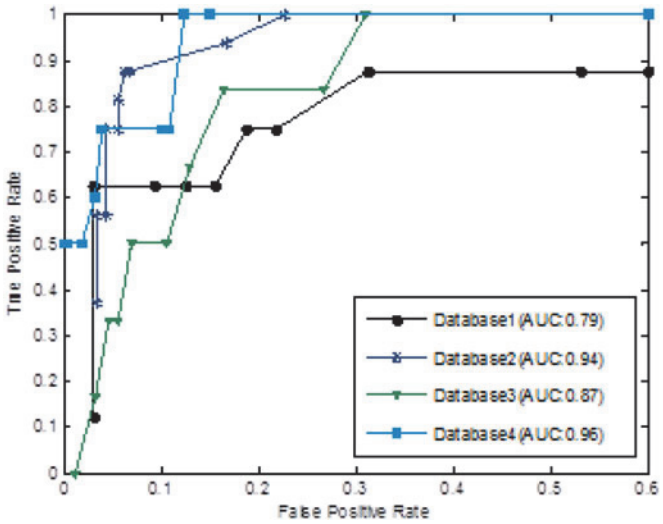

Fig. 6. ROC curves and the AUC values for the stationary datasets.

TABLE V
TRADITIONAL STATIONARY LEARNING OF COMPOSITE KERNELS

Datasets	SimpleMKL [24]	LMKL [20]	**LACK**
Stationary Set1	99.13	99.13	100
Stationary Set2	95.13	91.35	94.94
Stationary Set3	84.91	86.49	86.49
Stationary Set4	95.56	95.56	96.77

TABLE VI
EIGENVALUES λ OF 4 DIFFERENT KERNELS FOR NONSTATIONARY DATA SEQUENCES ♯1–♯10 FOR THE BASE NONSTATIONARY SETS ♯1–♯3

Datasets	Kernel	♯1	♯2	♯3	♯4	♯5	♯6	♯7	♯8	♯9	♯10
Non	Linear	18.9	8.43	25.7	11.8	31.6	14.9	27.8	19	29.3	16.8
stationary	Poly	18.3	8.41	24.8	9.59	28.5	13.2	24.1	14.4	28.5	16.2
Set1	RBF	20.5	8.76	28.7	9.43	30.1	16.3	25.5	14.1	27.9	13.3
	Laplace	11.7	7.4	24.5	11.4	27.9	14	23.2	17.8	21.8	14.8
Non	Linear	3.23	6.26	3.91	7.24	9.78	9.13	4.1	7.02	9.71	7.43
stationary	Poly	3.17	6.08	4.79	7.61	9.17	8.09	4.79	7.81	9.78	8.81
Set2	RBF	3.54	6.41	4.97	7.92	10.2	9.21	6.28	8.06	10.9	8.89
	Laplace	2.8	5.34	3.58	6.23	8.97	7.26	3.85	6.37	8.29	6.69
Non	Linear	21.3	34.8	18.8	22.1	25	35.9	24.1	24.1	24.1	27.8
stationary	Poly	33.7	43.6	19.3	22.8	26.1	37.1	25.8	27.9	28.6	29.6
Set3	RBF	34.1	45.7	20.3	24.7	28.9	38.6	28	28.7	29.1	30.6
	Laplace	21.8	20.9	17.4	20.8	18.3	29.7	21.8	21.5	16.8	23.1

generic kernel learning methods. It is also known that CAD is application specific, and thus, application of a generic machine learning algorithm to a CAD may not work well or achieve a satisfactory performance.

B. Anomaly Detection for the Nonstationary Data

We evaluated the proposed LACK in Section II to tune the selection of appropriate kernels when new nonstationary data become available. We divided the dataset equally into l sets (e.g., ten sets) to form the nonstationary data stream in Table VI, which shows the size of different nonstationary batches of data for each colon cancer dataset. Note that data sequences of l nonstationary subsets were generated from 3010 larger sequences with 600 anomaly and 2410 normal data sequences. Each nonstationary dataset was randomly divided into subsets. For example, the first sampled nonstationary dataset had 960 subsets; this was used for l small datasets (each dataset with 20 anomaly versus 76 normal data sequences). Others had another 1000 and 1050 in the total of three nonstationary datasets. After we tentatively formed the input matrices for four different kernels, we found the dominant kernels for the new data and the previous stationary data. These results are summarized in Table VI, in which the kernel with the maximum eigenvalue is highlighted for each nonstationary dataset in Table VI. We can see that the RBF kernel was always the most dominant kernel in all datasets, and the second dominant kernel kept varying.

TABLE VII
ANOMALY DETECTION WITH CLASSIFICATION ACCURACY AND AUC FOR NONSTATIONARY DATA SEQUENCES ♯1–♯10 FOR THE BASE NONSTATIONARY SETS ♯1–♯3

Datasets	Item	♯1	♯2	♯3	♯4	♯5	♯6	♯7	♯8	♯9	♯10
Non	Update	N	N	M	N	N	S	M	N	S	M
stationary	Accuracy	93.7	93.7	94.1	94.1	94.1	92.5	94.6	94.6	98.7	91.4
Set1	AUC	87.2	87.2	89.3	89.3	89.3	88.6	89.1	89.1	90.1	85.7
Non	Update	N	N	N	N	M	S	S	M	M	N
stationary	Accuracy	97.2	97.2	97.2	97.2	98.1	97.9	96.9	98	92.5	92.5
Set2	AUC	94.2	94.2	94.2	94.2	94.1	93.8	95	93.7	88.6	88.6
Non	Update	N	M	N	S	N	N	M	N	S	M
stationary	Accuracy	94.6	95.3	95.3	93.8	93.8	93.8	92.6	92.6	98.5	91.6
Set3	AUC	89.6	91.2	91.2	87.9	87.9	87.9	91.8	91.8	94.1	87.4

N indicates Normal, M indicates Minor Anomaly, and S indicates Significant Anomaly.

Table VII shows the results of detection of anomaly status using the proposed criterion of class separability $\xi = \lambda'_*/\lambda_*$ (5). In Section II, there were three statuses of the anomaly identification: normal (N), minor anomaly, (M), and significant anomaly (S), corresponding to the degree of anomaly. The classification performance was similar to stationary datasets. Each sequence for LACK was cascaded into one long sequence for the next experiment of large nonstationary sequential dataset. The anomaly detection performance was relatively high and is thus potentially useful for a preclinical setting of patients' diagnosis.

C. Time Horizontal Prediction for Risk Factor Analysis of Anomaly Long-Time Sequential Trajectories

We evaluated the predictive performance based on the anomaly trajectories, with the time horizontal transition of earlier low cancer stage (mostly called normal case) to the later higher cancer stage, using all long-time nonstationary sequential datasets in Table III. Note that, in this subsection, a large nonstationary dataset is much longer than the nonstationary datasets described previously. We adapted APH for assigning each cancer stage from 0 to III [3]. Using transition of the longitudinal sequential datasets, we applied anomaly detection to the entire nonstationary sequential datasets transient from normal to anomaly among cancer stage index 0–III. The posttest probability of APH was reconfigured to determine the chances that the patient had a disease. This synthesized measure incorporates the

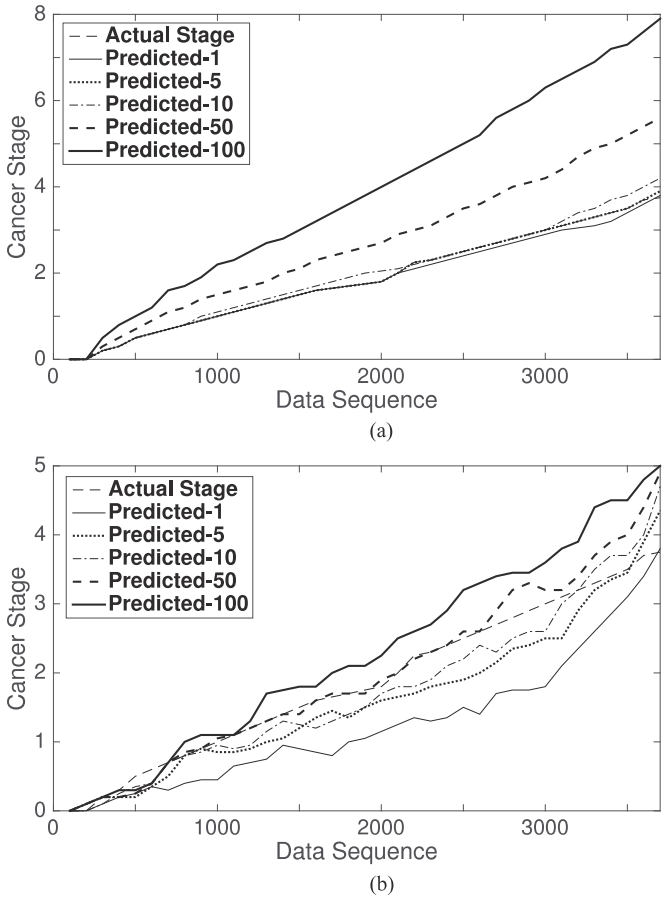

Fig. 7. Cancer stage of long-time sequence trajectories using (a) k-step prediction and (b) predictive LACK. When the horizontal time window increased 1, 5, 10, 50, 100 ahead of the time of data sequence, the predicted values were off from actual cancer stage value. (a) K-step and (b) predictive LACK performed much better than K-step under the larger prediction time frame, $k = 50$ or 100.

disease prevalence, the patient pool, specific patient risk factors (pretest odds), and information about the diagnostic test itself (the likelihood ratio) [6]. Fig. 7 shows the variable of time horizontal window, starting from each frame up to 3613 frames to analyze the risk of cancer stages. The representative size of the horizontal time window was set to predict cancer stages in advance. For example, in Fig. 7, these predicted window sizes k were $1, 5, 10, 50$, and 100 time index for the predicted values. The prediction window size $k = 1$ overlapped with the actual stage for both the k-step prediction method and the LACK prediction method. The k-step prediction of Fig. 7(a) shows larger error than LACK prediction of Fig. 7(b), especially when the predicted window size was 50 and 100.

These prediction results of Fig. 7 are summarized in Table VIII. The algorithms of k-step and the proposed LACK prediction were applied to all 3613 data sequences, which included a finite number of sequences called low cancer stage (I–II) and high cancer stage (II–III). For the comparison of prediction performance, a normalized metric was used, i.e., the NRMSE between the predicted cancer stage and the defined cancer stage. NRMSE is defined as a square root of the variance, known as the standard error between predicted cancer stage and actual ones. The metric is dimensionless so that it can make it possible to compare variable prediction accuracy with different time intervals. Fig. 7 and Table VIII show that NRMSE of II–III was larger than NRMSE of I–II. The traditional k-step and proposed LACK prediction methods were both efficient in handling nonstationary data over small window-size sequences with modest NRMSE. The prediction performance of the long prediction time window size $(50, 100)$ indicated that the prediction of the subsequent larger cancer stages of data using LACK was advantageous over a longer prediction horizontal window size. As shown in Table VIII, the proposed LACK outperformed the k-step prediction method as the horizontal time window size approached 100. This means average prediction error of LACK was relatively small even if the prediction was far ahead of time. The long-term prediction is, in general, preferable, since long-term analysis is clinically valuable to predict cancer stage.

TABLE VIII
AVERAGE ERRORS OF PREDICTION USING SEVERAL REPRESENTATIVE PREDICTION WINDOW SIZES

Horizontal Window Time Size	Low Cancer Stages (I–II)	High Cancer Stages (II–III)	Entire Cancer Stages
k-step 1	0.0029	0.006	0.0067
k-step 5	0.0294	0.0603	0.0671
k-step10	0.066	0.1356	0.1508
k-step 50	0.4125	0.8534	0.9478
k-step 100	1.0157	2.1205	2.3509
LACK step 1	0.1041	0.0904	0.098
LACK step 5	0.2777	0.4935	0.3934
LACK step 10	0.3348	0.686	0.5286
LACK step 50	0.5004	1.0387	0.7981
LACK step 100	0.7169	1.4919	1.1458

TABLE IX
COMPUTATIONAL TIME (MILLISECOND) OF EACH MODULE

Datasets	Kernel Selection	Cancer Classification	Normal Detection	Anomaly Prediction
Stationary Set1	5.7	0.81	NA	NA
Stationary Set2	31.9	4.1	NA	NA
Stationary Set3	43.2	4.9	NA	NA
Stationary Set4	63.4	6.1	NA	NA
Nonstationary Set1	19.8	2	0.62	0.61
Nonstationary Set2	22.3	2.1	0.71	0.72
Nonstationary Set3	30.1	2.4	1.09	1.08
Average	30.9	3.2	0.8	0.8

D. Computational Time for Complexity Evaluation

Table IX shows the computational cost of the proposed LACK over all four stationary and three nonstationary data without discarding it. In Table IX, the following four representative modules on LACK are listed: "Kernel Selection" indicates Section II for the results of Table IV, as well as "Cancer Classification" and "Anomaly Detection". "Anomaly Prediction" indicates Section III for the results of Fig. 7. We used a PC with an Intel i7 3.16-GHz CPU and 16-GB RAM. Table IX shows that, for the average of all seven different types of datasets, the

module of Kernel Selection along with stationary sets $2, 3, 4$ resulted in the highest computational load with an average of 30.9 ms among the four modules. The time reduction of nonstationary learning was 33.3% less than stationary learning. Thus, our proposed LACK computationally was very efficient for nonstationary learning. When adding anomaly detection and prediction, the overall computing cost did not change much due to the light modules.

V. Conclusion

This paper has proposed a novel method of detection and prediction of colon cancer staging from the long-time anomaly trajectories of nonstationary datasets using stationary training data. This method, called LACK, was designed to be faster and used a more efficient feature extraction algorithm than the previously developed KPCA method. The polyp classification experiment showed that LACK provided equivalent classification performance in nonstationary data compared to that of the stationary data. The k-step extension of LACK has the potential to predict the cancer staging index and yields high detection performance regarding the status of anomalies. Such an effective predictive scheme makes CT colonography a viable option for cancer staging analysis. The benefits of LACK include the ability to calculate the future likelihood of cancer stage progress and anomaly status, which may be deployed for the further diagnosis and treatment planning with a specific period and frequency.

References

[1] American Cancer Society, *Cancer Facts & Figures 2016*, Amer. Cancer Soc., Atlanta, GA, USA, 2016.

[2] B. Levin *et al.*, "Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: A joint guideline from the American Cancer Society," *CA Cancer J. Clin.*, vol. 58, pp. 130–160, 2008.

[3] J. Yee *et al.*, "ACR appropriateness criteria colorectal cancer screening," *J. Amer. Coll. Radiol.*, vol. 11, pp. 543–551, 2014.

[4] D. Regge and S. Halligan, "CAD: How it works, how to use it, performance," *Eur. J. Radiol.*, vol. 82, pp. 1171–1176, 2013.

[5] H. Yoshida, Y. Wu, and W. Cai, "Scalable, high-performance 3D image computing platform: System architecture and application to virtual colonoscopy," in *Proc. IEEE Eng. Med. Biol. Soc.*, 2012, pp. 3994–3997.

[6] A. Khan, J . A. Doucette, and R. Cohen, "Validation of an ontological medical decision support system for patient treatment using a repository of patient data: Insights into the value of machine learning," *ACM Trans. Intell. Syst. Technol.*, vol. 4, no. 4, 2013, Art. no. 68.

[7] P. Y. Chen, S. Yang, and J. A. McCann, "Distributed real-time anomaly detection in networked industrial sensing systems," *IEEE Trans. Ind. Electron.*, vol. 62, no. 6, pp. 3832–3842, Jun. 2015.

[8] L. Winter, Y. Motai, and A. Docef, "Computer-aided detection of polyps in CT colonography: On-line versus off-line accelerated kernel feature analysis," *Signal Process.*, vol. 90, pp. 2456–2467, 2010.

[9] H. Yoshida and J. Näppi, "CAD in CT colonography without and with oral contrast agents: Progress and challenges," *Comput. Med. Imag. Graph.*, vol. 31, pp. 267–84, 2007.

[10] B. J. Kim, I. K. Kim, and K. B. Kim, "Feature extraction and classification system for nonlinear and online data," *Adv. Knowl. Discovery Data Mining*, vol. 3056, pp. 171–180, 2004.

[11] G. Wang, J. Liu, and R. Srinivasan,"Data-driven soft sensor approach for quality prediction in a refining process," *IEEE Trans. Ind. Informat.*, vol. 6, no. 1, pp. 11–17, Feb. 2010.

[12] J.-H. Zhou, C. K. Pang, F. L. Lewis, and Z-.W. Zhong, "Intelligent diagnosis and prognosis of tool wear using dominant feature identification," *IEEE Trans. Ind. Informat.*, vol. 5, no. 4, pp. 454–464, Nov. 2009.

[13] X. Jiang, R. Snapp, Y. Motai, and X. Zhu, "Accelerated kernel feature analysis," *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recog.*, 2006, pp. 109–116.

[14] H. Xiong, M. N. S. Swamy, and M. O. Ahmad, "Optimizing the data-dependent kernel in the empirical feature space," *IEEE Trans. Neural Netw.*, vol. 16, no. 2, pp. 460–474, Mar. 2005.

[15] T. Briggs and T. Oates, "Discovering domain specific composite kernels," in *Proc. 20th Nat. Conf. Artif. Intell.*, 2005, pp. 732–738.

[16] A. Ning, H. C. W. Lau, Y. Zhao, and T. T. Wong, "Fulfillment of retailer demand by using the MDL-optimal neural network prediction and decision policy," *IEEE Trans. Ind. Informat.*, vol. 5, no. 4, pp. 495–506, Nov. 2009.

[17] W. Jiang, Z. Zhang, F. Li, L. Zhang, M. Zhao, and X. Jin, "Joint label consistent dictionary learning and adaptive label prediction for semisupervised machine fault classification," *IEEE Trans. Ind. Informat.*, vol. 12, no. 1, pp. 248–256, Feb. 2016.

[18] N. Cristianini, J. Kandola, A. Elisseeff, and J. Shawe-Taylor, "On kernel target alignment," *Proc. Neural Inf. Process. Syst. Conf.*, 2001, pp. 367–373.

[19] Y. Motai and H. Yoshida, "Principal composite kernel feature analysis: Data-dependent kernel approach," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 8, pp. 1863–1875, Aug. 2013.

[20] M. Gonen and E. Alpaydin, "Multiple kernel learning algorithms," *J. Mach. Learn. Res.*, vol. 12, pp. 2211–2268, 2011.

[21] S. Scardapane, D. Comminiello, M. Scarpiniti, and A. Uncini, "Online sequential extreme learning machine with kernels," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 9, pp. 2214–2220, Sep. 2015.

[22] O. Chapelle, V. Vapnik, O. Bousquest, and S. Mukherjee, "Choosing multiple parameters for support vector machines," *Mach. Learn.*, vol. 46, pp. 131–159, 2002.

[23] K. M. Chung, W. C. Kao, C.-L. Sun, L.-L. Wang, and C. J. Lin, "Radius margin bounds for support vector machines with the RBF kernel," *Neural Comput.*, vol. 15, no. 11, pp. 2643–2681, 2003.

[24] J. Yang, Z. Jin, J. Y. Yang, D. Zhang, and A. F. Frangi, "Essence of kernel Fisher discriminant: KPCA plus LDA," *Pattern Recog.*, vol. 37, pp. 2097–2100, 2004.

[25] M. O. Efe, O. Kaynak, and B. M. Wilamowski, "Stable training of computationally intelligent systems by using variable structure systems technique," *IEEE Trans. Ind. Electron.*, vol. 47, no. 2, pp. 487–496, Apr. 2000.

Yuichi Motai (M'01–SM'13) received the B.Eng. degree in instrumentation engineering from Keio University, Tokyo, Japan, in 1991, the M.Eng. degree in applied systems science from Kyoto University, Kyoto, Japan, in 1993, and the Ph.D. degree in electrical and computer engineering from Purdue University, West Lafayette, IN, USA, in 2002.

He is currently an Associate Professor of electrical and computer engineering with Virginia Commonwealth University, Richmond, VA, USA. His research interests include the broad area of sensory intelligence, particularly in medical imaging, pattern recognition, computer vision, and robotics.

Dingkun Ma received the B.Eng. degree in measuring and controlling technology from Zhengzhou University, Zhengzhou, China, in 2005, and the M.Eng. degree in mechatronical engineering and the Ph.D. degree in signal processing from Northwestern Polytechnical University, Xi'an, China, in 2008 and 2013, respectively.

He was a Visiting Research Scholar with Virginia Commonwealth University, Richmond, VA, USA, in 2010 and 2011.

Hiroyuki Yoshida (M'96) received the B.S. and M.S. degrees in physics and the Ph.D. degree in information science from the University of Tokyo, Tokyo, Japan, in 1983, 1985, and 1989 respectively.

He was an Assistant Professor in the Department of Radiology, University of Chicago. He was a Tenured Associate Professor when he left the university and joined the Massachusetts General Hospital (MGH) and Harvard Medical School (HMS), Boston, MA, USA, in 2005, where he is currently the Director of 3D Imaging Research in the Department of Radiology, MGH, and an Associate Professor of Radiology at HMS. His research interests include computer-aided diagnosis, in particular the diagnosis of polyps in computed tomographic colonography, for which he received several awards at the Annual Meetings of Radiological Society of North America and the International Society for Optical Engineering.